Reinforcement learning for human-robot shared control

Article (Accepted Version)

Reinforcement Learning for Human-Robot Shared Control

Yanan Li, *Member, IEEE*, Keng Peng Tee, *Member, IEEE*, Rui Yan, *Member, IEEE*, and Shuzhi Sam Ge, *Fellow, IEEE*

***Abstract*—This paper aims at proposing a general framework of shared control for human-robot interaction. Human dynamics are considered in analysis of the coupled human-robot system. Motion intentions of both human and robot are taken into account in the control objective of the robot. Reinforcement learning is developed to achieve the control objective subject to unknown dynamics of human and robot. The closed-loop system performance is discussed through a rigorous proof. Simulations are conducted to demonstrate the learning capability of the proposed method and its feasibility in handling various situations. Compared to existing works, the proposed framework combines motion intentions of both human and robot in a human-robot shared control system, without the requirement of the knowledge of humans and robots dynamics.**

I. INTRODUCTION

Despite the advent of robotics in the past several decades, the development of fully autonomous robots that fulfil operational requirements under real-world working conditions is still very challenging. The intervention of human beings is necessary in many complex tasks, especially when the working environment is unstructured and new to robots [1]. Researchers have seen the need for shared control of human and robot in extensive fields [2], [3], such as social applications [4], industrial settings [5], and space explorations [6], among others.

In the early literature of human-robot shared control, a leader-follower model is usually adopted wherein the robot is assigned a follower's role [7], [8]. This is mainly because most of current robots are still behind humans in the sense of intelligence. However, some researchers have recently recognized the importance of a framework beyond the conventional leader-follower model, such that human effort can be reduced and task performance improved by combining humans' and robots' advantages [9]. The initial efforts in this direction include recognition of human intention and evaluation of human performance, based on which a robot provides assistance (or, in relatively fewer cases, resistance) to the human partner whenever it is needed. In [10], human's motion intention is predicted based on a model acquired using probabilistic learning, and an optimal control framework is developed to simultaneously penalize the tracking error of the predicted motion and the control input of robot. This method is further improved in [11] by capturing the current unexpected human behaviors through online estimation of the current process noise. In [12], based on the objective of minimizing the human effort, an adaptation strategy is developed to switch between model-based and model-free predictions in the case of partially known tasks. Other works of intention prediction/estimation include: human's motion intention is represented by the change of the interaction force by assuming a preserved momentum [13]; the intentional walking direction of the users of a crane robot is estimated using a Kalman filter [14]; and the online estimation of human's motion intention is achieved based on the dynamic model of the human arm [15]. In [16], end-point impedance of human hands is measured to identify manual welding skills, of which the results can be used as clues of robotic assistance. In these works, a robot is expected to provide assistance to a human while it does not have its own objective. In many cases, however, a robot should also have its own objective since it can perform better than a human, e.g., a robot can precisely follow a predefined trajectory while the human partner intervenes it by moving it to several points-of-interest [17]. In this regard, we aim to develop a framework of human-robot shared control with both objectives of human and robot taken into account.

To study the human's control objective, it is essential to consider the human's dynamics which, however, are usually difficult to model. It raises such an issue that unknown human and robot dynamics will be involved in the system under study. To effectively cope with this issue, we will employ reinforcement learning in the control design. Reinforcement learning mimics the way that biological systems interact with environments [18]. It usually includes an actor that generates an action according to the environment stimuli, and a critic that evaluates the action result. Reinforcement learning has been extensively investigated in the early literature of machine learning [19], and its relationship to optimal control, learning control [20], [21], [22], [23] and adaptive control has gained recent attention of the control community [24]. Applications of reinforcement learning are found in cases when the exact system model cannot be easily obtained, such as missile systems [25] and power systems [26], etc. In theoretical developments, many research efforts have been

This work was supported by Grant No. 1225100001 from the Science and Engineering Research Council (SERC), A*STAR, Singapore.

Y. Li was with the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore 138632 and is with the Department of Engineering and Informatics, University of Sussex, Brighton, BN1 9RH, UK. yl557@sussex.ac.uk

K. P. Tee is with the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore 138632. kptee@i2r.a-star.edu.sg

R. Yan is with the College of Computer Science, Sichuan University, Chengdu 610065, China. ryan@scu.edu.cn

S. S. Ge is with the Department of Electrical and Computer Engineering and the Social Robotics Laboratory, Interactive Digital Media Institute, National University of Singapore, Singapore 117576. samge@nus.edu.sg

made on reinforcement learning control of continuous systems with completely unknown dynamics, which is still an open problem.

This paper will show a direct application of reinforcement learning, while a synchronized reinforcement learning control will be developed. Eventually, the control objective that combines both motion intentions of human and robot will be achieved in a human-robot shared control system, without the requirement of the knowledge of human's and robot's dynamics.

The rest of this paper is organized as follows. In Section II, the problem of human-robot shared control is formulated which includes the system description and the control objective. In Section III, the development of the proposed reinforcement learning control is detailed, in the presence of unknown human and robot dynamics. Simulation studies are designed to evaluate the performance of the proposed method, which will be discussed in Section IV. Concluding remarks are given in Section V.

II. PROBLEM FORMULATION

A. System Description

In this paper, we consider a general scenario where a human arm is in a physical interaction with a robot arm. There is a force/torque sensor at the interaction point on the robot arm so the interaction force (including force and torque) between the human and the robot can be measured. The robot arm accomplishes a task at its end-effector while the human arm applies an interaction force to influence the robot arm's movements. This scenario can be found in applications such as robotic welding [5] and object transporting [27].

The kinematic relationship of the robot arm can be described as

$$x(t) = \phi(q) \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is position/orientation in the Cartesian space, $q \in \mathbb{R}^n$ is coordinate in the joint space and $\phi(q) \in \mathbb{R}^{n\times n}$ is assumed to be nonsingular in a finite workspace. Differentiating the above equation with respect to time leads to

$$\dot{x}(t) = J(q)\dot{q} \tag{2}$$

where $J(q) \in \mathbb{R}^{n\times n}$ is the Jacobian matrix which is also assumed to be nonsingular in a finite workspace.

The dynamic model of the robot arm in the joint space is

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) = \tau + J^T(q)f(t) \tag{3}$$

where $M(q) \in \mathbb{R}^{n\times n}$ is the inertia matrix, $C(q,\dot{q})\dot{q} \in \mathbb{R}^n$ is the Coriolis and centrifugal force, $G(q) \in \mathbb{R}^n$ is the gravitational torque, $\tau \in \mathbb{R}^n$ is the control input, and $f(t) \in \mathbb{R}^n$ is the interaction force between the human and the robot in the Cartesian space.

Since the interaction happens in the Cartesian space, we substitute the kinematic constraints into the dynamic model (3), and obtain the dynamic model of the robot arm in the Cartesian space:

$$M_R(q)\ddot{x} + C_R(q,\dot{q})\dot{x} + G_R(q) = u + f(t) \tag{4}$$

where

$$\begin{aligned} &M_R(q) = J^{-T}(q)M(q)J^{-1}(q) \\ &C_R(q,\dot{q}) = J^{-T}(q)(C(q,\dot{q}) - M(q)J^{-1}(q)\dot{J}(q))J^{-1}(q) \\ &G_R(q) = J^{-T}(q)G(q),\ u = J^{-T}(q)\tau \end{aligned} \tag{5}$$

The above matrices and vectors have the following properties [28], which will be used in the control design.

Property 1: The matrix $2C_R - \dot{M}_R$ is a skew-symmetric matrix if C_R is in the Christoffel form, i.e., $\xi^T(2C_R - \dot{M}_R)\xi = 0,\ \forall \xi \in \mathbb{R}^n$.

Property 2: $b_M \le \|M_R\| \le b'_M$, $\|C_R\| \le b_C\|\dot{x}\|$ and $\|G_R\| \le b_G$, where b_M, b'_M, b_C and b_G are positive scalars.

The other part of the human-robot shared control system is the human arm. The following dynamic model of the human arm is employed, which is based on the hypothesis of equilibrium point control [29]:

$$C_H\dot{x} + K_H(x - x_H) = -f \tag{6}$$

where C_H and K_H are damping and stiffness matrices of the human arm, respectively, and x_H is the equilibrium position planned in the human's central nervous system (CNS). Eq. (6) is a simplified model with the inertia/mass component neglected, since it has been shown in [30] that the damping and stiffness components dominate the dynamics of the human arm. Note that C_H and K_H are time-varying because human may modulate the damping and stiffness of his/her arm in different stages of interaction. For analysis purpose, we assume that the damping matrix C_H is due to damping and Coriolis effects, so it is a function of x and $\dot{x}$. Referring to the stiffness ellipse in [29], the stiffness matrix K_H is position- and velocity-dependent, so it is also a function of x and $\dot{x}$. In summary, we have the following assumption:

Assumption 1: C_H and K_H are functions of x and $\dot{x}$.

From the above assumption, we can obtain the following lemma, which will be used in the control design:

Lemma 1: Given any vector $\xi \in \mathbb{R}^n$, we have

$$\begin{aligned} \dot{C}_H\xi &= C_{H1}(x,\xi)\dot{x} + C_{H2}(\dot{x},\xi)\ddot{x} \\ \dot{K}_H\xi &= K_{H1}(x,\xi)\dot{x} + K_{H2}(\dot{x},\xi)\ddot{x} \end{aligned} \tag{7}$$

where $C_{H1}(x,\xi)$, $C_{H2}(\dot{x},\xi)$, $K_{H1}(x,\xi)$, and $K_{H2}(\dot{x},\xi)$ are $n \times n$ matrices with the forms in the following proof.

Proof: Denote $\rho = \dot{C}_H\xi$, the vector composed by elements at the i-th row of C_H as C_i, and the element of C_H at the i-th row and j-th column as c_{ij}. According to Assumption 1, we have

$$\dot{c}_{ij} = \frac{\partial c_{ij}}{\partial x}\dot{x} + \frac{\partial c_{ij}}{\partial \dot{x}}\ddot{x} \tag{8}$$

Consider the i-th element of ρ, as below

$$\begin{aligned} \rho_i &= \Sigma_{j=1}^n \dot{c}_{ij}\xi_j \\ &= \Sigma_{j=1}^n[(\frac{\partial c_{ij}}{\partial x}\dot{x} + \frac{\partial c_{ij}}{\partial \dot{x}}\ddot{x})\xi_j] \\ &= \Sigma_{j=1}^n\Sigma_{k=1}^n[(\frac{\partial c_{ij}}{\partial x_k}\dot{x}_k + \frac{\partial c_{ij}}{\partial \dot{x}_k}\ddot{x}_k)\xi_j] \\ &= \Sigma_{k=1}^n(\frac{\partial C_i}{\partial x_k}\xi\dot{x}_k + \frac{\partial C_i}{\partial \dot{x}_k}\xi\ddot{x}_k) = \frac{\partial C_i}{\partial x^T}\xi\dot{x} + \frac{\partial C_i}{\partial \dot{x}^T}\xi\ddot{x} \end{aligned} \tag{9}$$

From the above equation, we find that the i-th rows of $C_{H1}(x,\xi)$ and $C_{H2}(\dot{x},\xi)$ are $\frac{\partial C_i}{\partial x^T}\xi$ and $\frac{\partial C_i}{\partial \dot{x}^T}\xi$, respectively. Therefore, we have

$$\dot{C}_H\xi = C_{H1}(x,\xi)\dot{x} + C_{H2}(\dot{x},\xi)\ddot{x} \tag{10}$$

Similarly, it can be shown that

$$\dot{K}_H\xi = K_{H1}(x,\xi)\dot{x} + K_{H2}(\dot{x},\xi)\ddot{x} \tag{11}$$

of which the details are omitted. □

B. Control Objective

As discussed in Introduction, the control framework under development is beyond the simple leader-follower model. In particular, we consider that both human and robot have their own objectives which may be coherent or conflicting. The overall control objective should be a trade-off of the two individual objectives. We describe it by introducing the following long-term discounted cost functional [31]:

$$\Pi(t) = \int_0^\infty e^{-\frac{s-t}{\psi}} r(x(s),u(s))\mathrm{d}s \tag{12}$$

where ψ is a time constant for discounting the future cost. $r(x(t),u(t))$ is the instant cost which is defined as

$$r = (x-x_d)^T Q_1 (x-x_d) + \dot{x}^T Q_2 \dot{x} + f^T Q_3 f + u^T R u \tag{13}$$

where x_d is the desired trajectory of the robot arm, $Q_1 \in \mathbb{R}^{n\times n} \geq 0$, $Q_2 \in \mathbb{R}^{n\times n} \geq 0$, and $Q_3 \in \mathbb{R}^{n\times n} \geq 0$ are the weights of position tracking, velocity regulation, and force regulation, respectively, and $R \in \mathbb{R}^{n\times n} > 0$ is the weight of the control input of the robot arm.

Remark 1: The last term of the instant cost r penalizes the control input of the robot arm. The first two terms penalize the error between the actual position (velocity) and the desired position (zero velocity) of the robot arm. Similarly, there should be a term to penalize the error between the actual position and the desired position of the human arm, i.e., $(x-x_H)$. However, x_H is unmeasurable, so it is replaced by the interaction force f. The rationale of the replacement can be understood from the dynamic model (6), which indicates that f is a measure of conflict between human intention and the actual position. By selecting different combinations of Q_1 and Q_3, we have different penalization of the robot's and human's objectives. For example, $Q_1 = 0$ and $Q_3 \neq 0$ indicate complete compliance to the human which corresponds to the traditional leader-follower framework. $Q_1 \neq 0$ and $Q_3 = 0$ indicate that the robot aims to track the desired trajectory and takes the interaction force as a disturbance, which corresponds to position control.

C. State-space Form

Since there are two parts, i.e., human and robot, in the system under study, we describe their dynamics in a unified form in this section. In particular, taking the derivative of the human dynamics (6) with respect to time leads to

$$\dot{C}_H\dot{x} + C_H\ddot{x} + \dot{K}_H(x-x_H) + K_H\dot{x} = -\dot{f} \tag{14}$$

Rearranging Eq. (14) and considering Eq. (6) and Assumption 1, we have

$$\begin{aligned}\dot{f} &= -K_{H1}(x,x-x_H)\dot{x} - K_{H2}(\dot{x},x-x_H)\ddot{x} - K_H\dot{x} \\ &\quad -C_{H1}(x,\dot{x})\dot{x} - C_{H2}(\dot{x})\ddot{x} - C_H\ddot{x}\end{aligned} \tag{15}$$

In the above equation, Lemma 1 is used where ξ is replaced by $x-x_H$ and $\dot{x}$. By omitting the arguments of K_{H1}, K_{H2}, C_{H1} and C_{H2}, the above equation is rewritten as

$$\dot{f} = -K_{H1}\dot{x} - K_{H2}\ddot{x} - K_H\dot{x} - C_{H1}\dot{x} - C_{H2}\ddot{x} - C_H\ddot{x} \tag{16}$$

Rearranging Eq. (4), we have

$$\begin{aligned}\ddot{x} &= -M_R^{-1}(q)C_R(q,\dot{q})\dot{x} - M_R^{-1}(q)G_R(q) + M_R^{-1}(q)f \\ &\quad +M_R^{-1}(q)u\end{aligned} \tag{17}$$

Choose three states $z_1 = x$, $z_2 = \dot{x}$, and $z_3 = f$ to form the system state $z = [z_1^T, z_2^T, z_3^T]^T$. Considering Eqs. (16) and (17), the system dynamics can be described as

$$\dot{z} = A(z)z + B(z)u + D(z) \tag{18}$$

where

$$\begin{aligned}A &= L^{-1}\begin{bmatrix}\mathbf{0}_n & I_n & \mathbf{0}_n \\ \mathbf{0}_n & -M_R^{-1}(q)C_R(q,\dot{q}) & M_R^{-1}(q) \\ \mathbf{0}_n & -K_{H1}-K_H-C_{H1} & \mathbf{0}_n\end{bmatrix} \\ B &= L^{-1}\begin{bmatrix}\mathbf{0}_n \\ M_R^{-1}(q) \\ \mathbf{0}_n\end{bmatrix},\ D = L^{-1}\begin{bmatrix}\mathbf{0}_n \\ -M_R^{-1}(q)G_R(q) \\ \mathbf{0}_n\end{bmatrix} \\ L &= \begin{bmatrix}I_n & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & I_n & \mathbf{0}_n \\ \mathbf{0}_n & C_{H2}+C_H+K_{H2} & I_n\end{bmatrix}\end{aligned} \tag{19}$$

$\mathbf{0}_n$ and I_n denote $n\times n$ zero and identity matrices, respectively. Considering $q = \phi^{-1}(x)$, $\dot{q} = J^{-1}\dot{x}$, $x - x_H = K_H^{-1}(f + C_H\dot{x})$, and Assumption 1, the argument of A, B, and D is z. Besides, it is trivial to prove that L^{-1} always exists by examining its eigenvalues.

In the following, we need further mathematical manipulation of the system model for the later control design. In particular, from Eq. (12) we see that the tracking error $x-x_d$ is in the cost functional, but it is not included in the state z. Therefore, this trajectory tracking problem needs to be transformed to a regulation problem for the employment of reinforcement learning. To this end, the desired trajectory of the robot arm x_d is assumed to be generated by the following system

$$\dot{x}_d = F(x_d) \tag{20}$$

where $F(\cdot)$ is a function given by the designer. It can be either linear or nonlinear but known to the robot. Then, we consider the augmented state $\bar{z} = [z^T, x_d^T]^T$. By combining Eqs. (18) and (20), we have the augmented system

$$\dot{\bar{z}} = \bar{A}(\bar{z}) + \bar{B}(\bar{z})u \tag{21}$$

where

$$\bar{A}(\bar{z}) = \begin{bmatrix}A(z)z + D(z) \\ F(x_d)\end{bmatrix},\ \bar{B}(\bar{z}) = \begin{bmatrix}B(z) \\ \mathbf{0}_n\end{bmatrix} \tag{22}$$

Then, the cost functional (12) can be rewritten as

$$\Pi = \int_0^\infty e^{-\frac{s-t}{\psi}} [\bar{z}^T(s) Q \bar{z}(s) + u^T(s) R u(s)] \mathrm{d}s \tag{23}$$

where

$$Q = \begin{bmatrix} Q_1 & \mathbf{0}_n & \mathbf{0}_n & -Q_1 \\ \mathbf{0}_n & Q_2 & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n & Q_3 & \mathbf{0}_n \\ -Q_1 & \mathbf{0}_n & \mathbf{0}_n & Q_1 \end{bmatrix} \tag{24}$$

It is easy to verify that $Q \geq 0$.

Now we arrive at the statement of control objective, which is to design u to minimize the cost functional (23) by considering the system dynamics (21). It is clear that $\bar{A}(\bar{z})$, $\bar{B}(\bar{z})$ and $\bar{D}(\bar{z})$ are usually unknown or uncertain because unknown and uncertain human and robot dynamics are involved. Therefore, we will design a reinforcement learning control method in the remainder of this paper, in order to cope with this issue.

D. Preliminary: Function Approximation

Function approximation will be needed in the control design, and it has been extensively studied in the literature [32]. Among these existing approaches, we employ radial basis function neural network (RBFNN) in this paper, which has been widely used to estimate nonlinear functions due to its capabilities in function approximation. It has been shown that RBFNN is able to estimate any continuous function $h(Z): R^p \to R^y$ over a compact set $\Omega_Z \subset R^p$ to any arbitrary accuracy [33], i.e.,

$$h(Z) = W^T S(Z) + \epsilon, \ \forall Z \in \Omega_Z \tag{25}$$

where the input vector $Z \in \Omega \subset R^p$, $W \in R^{l \times y}$ is the ideal constant weight and l is the number of nodes which is greater than 1, and ϵ is the approximation error. $S(Z) = [s_1(Z), s_2(Z), ..., s_l(Z)]^T$ where $s_i(Z)$ can be a Gaussian function as below

$$s_i(Z) = \exp\Big[\frac{-(Z-\mu_i)^T(Z-\mu_i)}{\eta_i^2}\Big] \tag{26}$$

where $i = 1, 2, ..., l, \mu_i = [\mu_{i1}, \mu_{i2}, ..., \mu_{ip}]$ is the center of receptive field, and η_i is the width of the Gaussian function.

III. REINFORCEMENT LEARNING

A. Control Design

The control design follows the design of a typical reinforcement learning control, which includes two parts: critic and actor. In particular, a critic network is developed to evaluate the action result, while an actor network is developed to generate an action to minimize a predefined cost functional. They are detailed in the following two subsections, respectively.

1) Critic Network: Denote the value function associated with the cost functional Π in (23) as

$$\Gamma = \int_t^\infty e^{-\frac{s-t}{\psi}} r(x(s), u(s)) \mathrm{d}s \tag{27}$$

which is the solution of the Hamilton-Jacobian-Bellman (HJB) equation:

$$\begin{aligned} 0 \ &= \ \bar{z}^T Q \bar{z} + (\nabla\Gamma)^T \bar{A}(\bar{z}) \\ &\quad -\frac{1}{4}(\nabla\Gamma)^T \bar{B}(\bar{z}) R^{-1} \bar{B}^T(\bar{z}) \nabla\Gamma \end{aligned} \tag{28}$$

where $\nabla\Gamma = \frac{\partial\Gamma}{\partial\bar{z}}$. Since it is difficult to obtain Γ by solving the above equation, we introduce a critic network to approximate it, i.e.,

$$\Gamma = W_c^T S_c(\bar{z}) + \epsilon_c, \ \hat{\Gamma} = \hat{W}_c^T S_c(\bar{z}) \tag{29}$$

where denotations follow the definition of NN in Section II-D and $_c$ stands for critic. In particular, W_c is the ideal constant weight of the critic network, $S_c(\bar{z})$ the activation function, and ϵ_c the approximation error. From the definition (27), the estimation error of the cost-to-go function is [31]:

$$e_c = r - \frac{1}{\psi}\hat{\Gamma} + \hat{W}_c^T \nabla S_c \dot{\bar{z}} \tag{30}$$

where $\nabla S_c = \frac{\partial S_c}{\partial \bar{z}}$.

Define an error function as below

$$E_c = \frac{1}{2} e_c^2 \tag{31}$$

Then, the update law of the critic network can be designed as

$$\dot{\hat{W}}_c' = -\sigma_c \frac{\partial E_c}{\partial \hat{W}_c'} \tag{32}$$

where $\sigma_c > 0$ is the learning rate for the critic network. Considering Eq. (30), we have

$$\dot{\hat{W}}_c' = -\sigma_c e_c (-\frac{1}{\psi}\frac{\partial\hat{\Gamma}}{\partial\hat{W}_c'} + \nabla S_c \dot{\bar{z}}) \tag{33}$$

However, since unknown dynamics $\dot{\bar{z}}$ (more exactly, $\dot{z}$ in Eq. (21)) are involved in the above update law, we need to develop an identifier to estimate them [34]. In particular, we rewrite Eq. (21) in the form of NN approximation as below

$$\dot{z} = W_{id}^T S_{id}(z, u) + \epsilon_{id} \tag{34}$$

where $_{id}$ stands for identification. Then, $\dot{z}$ can be approximated by

$$\dot{\hat{z}} = \hat{W}_{id}^T S_{id}(\hat{z}, u) - K_{id}\tilde{z} \tag{35}$$

where $\tilde{z} = \hat{z} - z$ and $K_{id} > 0$. Similarly, denotations of the NN approximation follow those in Section II-D.

Denote $W_{id,i}$ and $\hat{W}_{id,i}$ as the ith columns of W_{id} and $\hat{W}_{id}$, respectively, and z_i, $\hat{z}_i$, and $\tilde{z}_i$ the ith components of z, $\hat{z}$, and $\tilde{z}$, respectively, $i = 1, 2, \ldots, 3n$. Then, the update law of $\hat{W}_{id}$ is given as

$$\dot{\hat{W}}_{id,i} = -S_{id}\tilde{z}_i - \sigma_{id}\hat{W}_{id,i} \tag{36}$$

where $\sigma_{id} > 0$. Define $\dot{\bar{\hat{z}}} = [\dot{\hat{z}}^T, \dot{x}_d^T]^T$ and

$$\Lambda = -\frac{1}{\psi} S_c(\bar{z}) + \nabla S_c \dot{\bar{\hat{z}}} \tag{37}$$

Then, the update law of the critic network becomes

$$\dot{\hat{W}}_c = -\sigma_c(r(t) + \hat{W}_c^T\Lambda)\Lambda \tag{38}$$

Remark 2: Instead of a system identifier, integral reinforcement learning (IRL) in [34] can be also employed to avoid the requirement of the knowledge of the system dynamics, which has been extended for the optimal tracking control problem of continuous-time nonlinear systems in [35].

2) Actor Network: Now, we introduce an actor network to obtain the control input. Define the tracking error

$$e = x - x_d \tag{39}$$

and consider a Lyapunov function candidate

$$V_1 = \frac{1}{2}e^Te \tag{40}$$

The derivative of V_1 with respect to time is

$$\dot{V}_1 = e^T\dot{e} = e^T(\dot{x} - \dot{x}_d + K_1e - K_1e) \tag{41}$$

where K_1 is a positive definite matrix. Then, define

$$\begin{aligned} \dot{x}_r &= \dot{x}_d - K_1e \\ e_v &= \dot{x} - \dot{x}_r \end{aligned} \tag{42}$$

The derivative of V_1 becomes

$$\dot{V}_1 = -e^TK_1e + e^Te_v \tag{43}$$

Considering the definition of e_v and Eq. (17), we have

$$\dot{e}_v = -M_R^{-1}C_R\dot{x} - M_R^{-1}G_R + M_R^{-1}f + M_R^{-1}u - \ddot{x}_r \tag{44}$$

where $\ddot{x}_r = \ddot{x}_d - K_1\dot{e}$.

Consider another Lyapunov function candidate

$$V_2 = \frac{1}{2}e_v^TM_Re_v \tag{45}$$

The derivative of V_2 with respect to time is

$$\begin{aligned} \dot{V}_2 &= \frac{1}{2}e_v^T\dot{M}_Re_v + e_v^TM_R\dot{e}_v \\ &= \frac{1}{2}e_v^T\dot{M}_Re_v - e_v^T(C_R\dot{x} + G_R - f - u + M_R\ddot{x}_r) \end{aligned} \tag{46}$$

Applying Property 1, we have

$$\dot{V}_2 = -e_v^T(C_R\dot{x}_r + G_R - f - u + M_R\ddot{x}_r) \tag{47}$$

We employ an actor network to estimate the desired control

$$u_d = W_a^TS_a(Z_a) + \epsilon_a - f - e - K_2e_v \tag{48}$$

where K_2 is a positive definite matrix, $_d$ stands for desired, $_a$ stands for actor, and

$$Z_a = [q, \dot{q}, \dot{x}_r, \ddot{x}_r] \tag{49}$$

The actual control with the estimated weight is thus given by

$$u = \hat{W}_a^TS_a(Z_a) - f - e - K_2e_v \tag{50}$$

Considering Eqs. (43) and (47), we have

$$\dot{V}_1 + \dot{V}_2 = -e^TK_1e - e_v^TK_2e_v + e_v^T(\tilde{W}_a^TS_a - \epsilon_a) \tag{51}$$

where $\tilde{W}_a = \hat{W}_a - W_a$. The objective of the update law is to make the estimation error of the actor network and value function converge, so we define

$$e_a = \sum_{i=1}^{n}\tilde{W}_{a,i}^TS_a + k_\Gamma\hat{\Gamma},\ E_a = \frac{1}{2}e_a^2 \tag{52}$$

where $\tilde{W}_{a,i}$ is the ith column of $\tilde{W}_a$ for $i = 1, 2, \ldots, n$, and k_Γ is a positive constant. Therefore, the update law can be designed based on gradient descent, i.e.,

$$\dot{\hat{W}}'_{a,i} = -\sigma_a\frac{\partial E_a}{\partial \hat{W}'_{a,i}} = -\sigma_a(\sum_{i=1}^{n}\tilde{W}_{a,i}^TS_a + k_\Gamma\hat{\Gamma})S_a \tag{53}$$

where $\sigma_a > 0$ is the learning rate for the actor network. However, since $\tilde{W}_{a,i}$ is unknown, the following update law is developed

$$\dot{\hat{W}}_{a,i} = -\sigma_a(\hat{W}_{a,i}^TS_a + k_\Gamma\hat{\Gamma})S_a \tag{54}$$

Remark 3: The proposed control (50) is a synchronized reinforcement learning control in the sense that its weight $\hat{W}_a$ is updated in (54) simultaneously with the weight of the critic network $\hat{W}_c$ updated in (38), without the requirement of the knowledge of the system dynamics.

B. Stability Analysis

Assumption 2: [36] Let the signals Λ and S_a be persistently exciting (PE) over the time interval $[t, t+T]$, i.e., there exist constants $b_1 > 0$, $b_2 > 0$, $b_3 > 0$ and $b_4 > 0$ such that

$$\begin{aligned} b_1I_{l_c} \le \int_t^{t+T}\Lambda(s)\Lambda^T(s)\mathrm{d}s \le b_2I_{l_c} \\ b_3I_{l_a} \le \int_t^{t+T}S_a(s)S_a^T(s)\mathrm{d}s \le b_4I_{l_a} \end{aligned} \tag{55}$$

Theorem 1: Consider robot dynamics (4) and human dynamics (6) with Assumption 1. The proposed control (50), with update laws (38) and (54), guarantees that the closed-loop system signals e, $\tilde{W}_c$, and $\tilde{W}_a$ converge to a compact set $\Omega := \left\{\omega |\ \|\omega\| \le \sqrt{\chi}\right\}$, where $\chi = \frac{b}{\kappa}$ and

$$\begin{aligned} \kappa = \min(&\|K_1 - \sigma_cc_1\varepsilon_{c2}^2I_n\|, \|K_2 - (1 + \sigma_cc_2\varepsilon_{c2}^2)I_n\|, \frac{b_\Lambda}{T}, \\ &\|K_{id} - \frac{1}{2}I_{3n}\|, \sigma_{id} - 2\sigma_cb_\nabla^2, \frac{b_S}{T}) \\ b = &\frac{b_{id}^2}{2} + 2\sigma_cb_\nabla^2b_{id}^2 + \frac{\varepsilon_a^2}{2} + \sigma_ab_a^2 + 2n\sigma_ak_\Gamma^2b_c^2 \\ &+\sigma_c(\varepsilon_{c1}^2 + c_4\varepsilon_{c2}^2) \end{aligned}$$

where $b_\nabla \ge \|W_c^T\nabla S_c\|$, $b_{id} \ge \|W_{id}^T(S_{id}(\hat{z}, u) - S_{id}(z, u)) - \epsilon_{id}\|$, $\varepsilon_a \ge \|\epsilon_a\|$, $b_a \ge \|W_a\|$, $b_c \ge \|W_c\|$, $\varepsilon_{c1} \ge \|\frac{\epsilon_c}{\psi}\|$ and $\varepsilon_{c2} \ge \|\frac{\partial\epsilon_c}{\partial\bar{z}}\|$. b_Λ, b_S and c_i, $i = 1, 2, 3, 4$ will be defined in the following proof. Moreover, the following conditions must be fulfilled:

$$\begin{aligned} &K_1 - \sigma_cc_1\varepsilon_{c2}^2I_n > 0,\ K_2 - (1 + \sigma_cc_2\varepsilon_{c2}^2)I_n > 0, \\ &K_{id} - \frac{1}{2}I_{3n} > 0,\ \sigma_{id} - 2\sigma_cb_\nabla^2 > 0 \end{aligned}$$

Proof: Consider a Lyapunov function candidate as below

$$\begin{aligned} V &= V_1 + V_2 + V_{id} + V_c + V_a \\ V_{id} &= \frac{1}{2}\tilde{z}^T\tilde{z} + \frac{1}{2}\sum_{i=1}^{3n}\tilde{W}_{id,i}^T\tilde{W}_{id,i} \\ V_c &= \frac{1}{2}\tilde{W}_c^T\tilde{W}_c,\ V_a = \frac{1}{2}\sum_{i=1}^{n}\tilde{W}_{a,i}^T\tilde{W}_{a,i} \end{aligned} \tag{56}$$

where $\tilde{W}_{id,i} = \hat{W}_{id,i} - W_{id,i}$, $i = 1, 2, \ldots, 3n$, and $\tilde{W}_c = \hat{W}_c - W_c$.

Subtracting (34) by (35), we have

$$\begin{aligned} \dot{\tilde{z}} &= \tilde{W}_{id}^T S_{id}(\hat{z}, u) - K_{id}\tilde{z} + W_{id}^T(S_{id}(\hat{z}, u) - S_{id}(z, u)) \\ &\quad -\epsilon_{id} \end{aligned} \tag{57}$$

Considering Eqs. (36) and (57), we have

$$\begin{aligned} \dot{V}_{id} &= -\tilde{z}^T K_{id}\tilde{z} - \sigma_{id}\sum_{i=1}^{3n}\tilde{W}_{id,i}^T\tilde{W}_{id,i} \\ &\quad +\tilde{z}^T(W_{id}^T(S_{id}(\hat{z}, u) - S_{id}(z, u)) - \epsilon_{id}) \\ &\leq -\tilde{z}^T(K_{id} - \frac{1}{2}I_{3n})\tilde{z} - \sigma_{id}\sum_{i=1}^{3n}\tilde{W}_{id,i}^T\tilde{W}_{id,i} + \frac{1}{2}b_{id}^2 \end{aligned} \tag{58}$$

According to Eq. (38), the derivative of V_c is

$$\dot{V}_c = -\sigma_c\tilde{W}_c^T(r + \hat{W}_c^T\Lambda)\Lambda \tag{59}$$

Differentiating Eq. (27) leads to

$$r = \frac{1}{\psi}\Gamma - \dot{\Gamma} \tag{60}$$

Thus, we obtain

$$\begin{aligned} r &= \frac{1}{\psi}(W_c^T S_c(\bar{z}) + \epsilon_c) - W_c^T\nabla S_c\dot{\tilde{z}} - \dot{\epsilon}_c \\ &= -W_c^T\Lambda + W_c^T\nabla S_c\dot{\tilde{z}} + (\frac{\epsilon_c}{\psi} - \dot{\epsilon}_c) \end{aligned} \tag{61}$$

Substituting Eq. (61) into Eq. (59), we obtain

$$\begin{aligned} \dot{V}_c &\leq -\frac{\sigma_c}{2}\tilde{W}_c^T\Lambda\Lambda^T\tilde{W}_c + \frac{\sigma_c}{2}(W_c^T\nabla S_c\dot{\tilde{z}} + (\frac{\epsilon_c}{\psi} - \dot{\epsilon}_c))^2 \\ &\leq -\frac{\sigma_c}{2}\tilde{W}_c^T\Lambda\Lambda^T\tilde{W}_c + \sigma_c\varepsilon_{c1}^2 + \sigma_c\|\dot{\epsilon}_c\|^2 + \sigma_c b_\nabla^2\|\dot{\tilde{z}}\|^2 \end{aligned} \tag{62}$$

Substituting Eq. (57) to Ineq. (62), we have

$$\begin{aligned} \dot{V}_c &\leq -\frac{\sigma_c}{2}\tilde{W}_c^T\Lambda\Lambda^T\tilde{W}_c + 2\sigma_c b_\nabla^2\sum_{i=1}^{3n}\tilde{W}_{id,i}^T\tilde{W}_{id,i} \\ &\quad +2\sigma_c b_\nabla^2 b_{id}^2 + \sigma_c\varepsilon_{c1}^2 + \sigma_c\|\dot{\epsilon}_c\|^2 \end{aligned} \tag{63}$$

By assuming $\|\dot{f}\| \leq b_f$ and $\|\dot{x}_d\| \leq b_{xd}$, considering Property 2, and recalling $\dot{\tilde{z}} = [\dot{x}^T, \ddot{x}^T, \dot{f}^T, \dot{x}_d]^T$, $\ddot{x}$ in Eq. (17), u in Eq. (50) and

$$\dot{x} = e_v + \dot{x} - K_1 e \tag{64}$$

we have

$$\|\dot{\tilde{z}}\|^2 \leq (c_1\|e\|^2 + c_2\|e_v\|^2 + c_3\|\tilde{W}_a\|^2 + c_4) \tag{65}$$

where $c_i,\ i = 1, 2, 3, 4$ are positive constants. Furthermore, we obtain

$$\|\dot{\epsilon}_c\|^2 \leq \varepsilon_{c2}^2(c_1\|e\|^2 + c_2\|e_v\|^2 + c_3\|\tilde{W}_a\|^2 + c_4) \tag{66}$$

The derivative of V_a with respect to time is

$$\begin{aligned} \dot{V}_a &= -\sigma_a\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a(\hat{W}_{a,i}^T S_a + k_\Gamma\hat{\Gamma}) \\ &= -\sigma_a\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a S_a^T\tilde{W}_{a,i} - \sigma_a\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a(W_{a,i}^T S_a + k_\Gamma\hat{\Gamma}) \\ &\leq -\frac{\sigma_a}{2}\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a S_a^T\tilde{W}_{a,i} + \frac{\sigma_a}{2}\sum_{i=1}^{n}(W_{a,i}^T S_a + k_\Gamma\hat{\Gamma})^2 \\ &\leq -\frac{\sigma_a}{2}\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a S_a^T\tilde{W}_{a,i} + \sigma_a\sum_{i=1}^{n}(S_a^T S_a W_{a,i}^T W_{a,i} + k_\Gamma^2\hat{\Gamma}^2) \\ &\leq -\frac{\sigma_a}{2}\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a S_a^T\tilde{W}_{a,i} \\ &\quad +\sigma_a\sum_{i=1}^{n}\left(W_{a,i}^T W_{a,i} + 2k_\Gamma^2(\tilde{W}_c^T\tilde{W}_c + W_c^T W_c)\right) \end{aligned} \tag{67}$$

Considering (43), (47), (63), (66) and (67), we obtain

$$\begin{aligned} \dot{V} &\leq -e^T K_1 e - e_v^T K_2 e_v + e_v^T(\tilde{W}_a^T S_a - \epsilon_a) \\ &\quad -\frac{\sigma_a}{2}\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a S_a^T\tilde{W}_{a,i} \\ &\quad +\sigma_a\sum_{i=1}^{n}\left(W_{a,i}^T W_{a,i} + 2k_\Gamma^2(\tilde{W}_c^T\tilde{W}_c + W_c^T W_c)\right) \\ &\quad -\tilde{z}^T(K_{id} - \frac{1}{2}I_{3n})\tilde{z} + \frac{b_{id}^2}{2} - \frac{\sigma_c}{2}\tilde{W}_c^T\Lambda\Lambda^T\tilde{W}_c \\ &\quad -(\sigma_{id} - 2\sigma_c b_\nabla^2)\sum_{i=1}^{3n}\tilde{W}_{id,i}^T\tilde{W}_{id,i} + 2\sigma_c b_\nabla^2 b_{id}^2 \\ &\quad +\sigma_c\varepsilon_{c1}^2 + \sigma_c\|\dot{\epsilon}_c\|^2 \end{aligned} \tag{68}$$

Consider the following inequalities

$$\begin{aligned} -e_v^T\epsilon_a &\leq \frac{\|e_v\|^2}{2} + \frac{\|\epsilon_a\|^2}{2} \\ e_v^T\tilde{W}_a^T S_a &\leq \frac{\|e_v\|^2}{2} + \frac{\|\tilde{W}_a^T S_a\|^2}{2} \leq \frac{\|e_v\|^2}{2} + \sum_{i=1}^{n}\tilde{W}_{a,i}^T\tilde{W}_{a,i} \end{aligned}$$

Substituting them into Ineq. (68), we have

$$\begin{aligned} \dot{V} &\leq -e^T(K_1 - \sigma_c c_1\varepsilon_{c2}^2 I_n)e - e_v^T(K_2 - (1 + \sigma_c c_2\varepsilon_{c2}^2)I_n)e_v \\ &\quad -\tilde{z}^T(K_{id} - \frac{1}{2}I_{3n})\tilde{z} - (\sigma_{id} - 2\sigma_c b_\nabla^2)\sum_{i=1}^{3n}\tilde{W}_{id,i}^T\tilde{W}_{id,i} \\ &\quad -\frac{\sigma_c}{2}\tilde{W}_c^T\Lambda\Lambda^T\tilde{W}_c + 2n\sigma_a k_\Gamma^2\tilde{W}_c^T\tilde{W}_c \\ &\quad -\frac{\sigma_a}{2}\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a S_a^T\tilde{W}_{a,i} + (1 + \sigma_c c_3\varepsilon_{c2}^2)\sum_{i=1}^{n}\tilde{W}_{a,i}^T\tilde{W}_{a,i} \\ &\quad +\frac{b_{id}^2}{2} + 2\sigma_c b_\nabla^2 b_{id}^2 + \frac{\|\epsilon_a\|^2}{2} + \sigma_a\|W_a\|^2 \\ &\quad +2n\sigma_a k_\Gamma^2\|W_c\|^2 + \sigma_c(\varepsilon_{c1}^2 + c_4\varepsilon_{c2}^2) \end{aligned} \tag{69}$$

Therefore, for $t > t_0$ (t_0 is a certain time), $\|e\|^2 \leq \frac{b}{\|K_1 - \sigma_c c_1\varepsilon_{c2}^2 I_n\|}$, $\|e_v\|^2 \leq \frac{b}{\|K_2 - (1+\sigma_c c_2\varepsilon_{c2}^2)I_n\|}$, $\|\tilde{z}\|^2 \leq$

$\frac{b}{\|K_{id}-\frac{1}{2}I_{3n}\|}$, $\sum_{i=1}^{3n}\tilde{W}_{id,i}^T\tilde{W}_{id,i} \leq \frac{b}{\|\sigma_{id}-2\sigma_c b_\nabla^2\|}$, and

$$\frac{\sigma_c}{2}\tilde{W}_c^T\Lambda\Lambda^T\tilde{W}_c - 2n\sigma_a k_\Gamma^2\tilde{W}_c^T\tilde{W}_c \leq b$$

$$\frac{\sigma_a}{2}\sum_{i=1}^{n}\tilde{W}_{a,i}^T S_a S_a^T\tilde{W}_{a,i} - (1+\sigma_c c_3\varepsilon_{c2}^2)\sum_{i=1}^{n}\tilde{W}_{a,i}^T\tilde{W}_{a,i} \leq b \quad (70)$$

Considering the update laws (38) and (54), we have

$$\dot{\tilde{W}}_c = -\sigma_c\Lambda\Lambda^T\tilde{W}_c^T - \sigma_c\Lambda(W_c^T\nabla S_c\dot{\tilde{z}} + (\frac{\epsilon_c}{\psi} - \dot{\epsilon}_c)) \quad (71)$$

$$\dot{\tilde{W}}_{a,i} = -\sigma_a S_a S_a^T\tilde{W}_{a,i} - \sigma_a S_a(W_{a,i}^T S_a + k_\Gamma\hat{\Gamma}) \quad (72)$$

According to Technical Lemma 1 in [37] and Assumption 1, we have

$$b_\Lambda I_{l_c} \leq \int_t^{t+T}\psi_c^T(s,t)(\frac{\sigma_c}{2}\Lambda(s)\Lambda^T(s) - 2n\sigma_a k_\Gamma^2 I_{l_c}) \times\psi_c(s,t)\mathrm{d}s \leq b'_\Lambda I_{l_c}$$

$$b_S I_{l_a} \leq \int_t^{t+T}\psi_a^T(s,t)(\frac{\sigma_a}{2}S_a S_a^T - (1+\sigma_c c_3\varepsilon_{c2}^2)I_{l_a}) \times\psi_a(s,t)\mathrm{d}s \leq b'_S I_{l_a}$$

where $b_\Lambda > 0$, $b'_\Lambda > 0$, $b_S > 0$ and $b'_S > 0$, and ψ_c and ψ_a are the state transition matrices of (71) and (72), respectively. Note that we use a fact that the unit matrices I_{l_c} and I_{l_a} satisfy the PE condition. Integrating both sides of Ineq. (70) from t to $t+T$, we have $b_\Lambda\|\tilde{W}_c\|^2 \leq bT$ and $b_S\|\tilde{W}_a\|^2 \leq bT$, which completes the proof. □

Remark 4: In [38], [24], an actor network in the form of $u = -\frac{1}{2}R^{-1}\bar{B}^T(\bar{z})\Gamma_{\bar{z}}$ is developed to ensure convergence to the optimal control. In this paper, we develop an actor network as in Eq. (50) with Eq. (54). According to Theorem 1, $\dot{\hat{W}}_{a,i}$ in Eq. (54) can be very small and $\hat{W}_{a,i}$ can be very near to $W_{a,i}$ under the proposed actor network. By choosing a large enough k_Γ, the approximated value function $\hat{\Gamma}$ can be also very small, although the optimal solution is not achieved. The same technique can be found in [39]. However, selection of a large k_Γ results in a large bound b which may cause instability. In this regard, it needs to be properly chosen to achieve a good balance between control performance and stability.

Remark 5: In the literature, many methods are proposed to address the exploration issue in reinforcement learning, such as state resetting and covariance resetting [40]. Among them, injection of an exploration noise into the control input is employed in above stability analysis to satisfy the persistent excitation condition as efficient exploration, similarly as in [34], [35]. It is generally nontrivial to choose the exploration noise and there is a dilemma between the efficient exploration and the satisfactory control performance [41]. In the human-in-the-loop system under study in this paper, an improper choice of the exploration noise may even cause disturbance to the human. Therefore, a good balance between the exploration and exploitation should be found through trial and error.

IV. SIMULATION

The simulation scenario is sketched in Fig. 1, where a human hand holds the end-effector of a planar robot arm with two revolute joints. A desired trajectory is prescribed for the robot arm and there are human's areas of interest which may or may not include the desired trajectory of the robot arm. In this case, human will move the robot arm to these areas with equilibrium positions of his/her arm, which are illustrated by pentagrams. In this scenario, different portions of human and robot controls are required due to different distances between the desired trajectory of the robot arm and human's area of interest: if human's area of interest is near to the desired trajectory of the robot arm, a small (large) portion of human (robot) control is needed, and vice versa. This is different from the model where either the human or robot takes full control in a switching manner, which is undesirable because full human control unnecessarily requires more human effort.

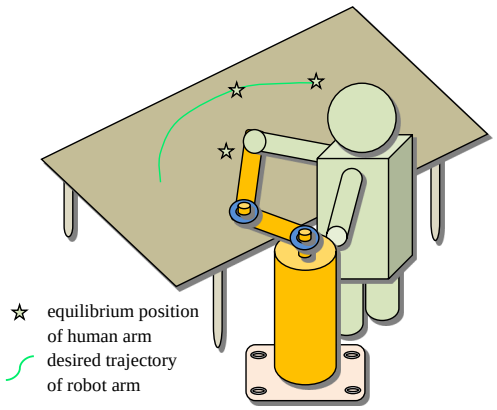

Fig. 1. Simulation scenario: the robot arm has a prescribed desired trajectory while the human tries to move it to the equilibrium positions of his/her arm

The desired trajectory of the robot arm is half a circle with a radius of 0.2m, from $[-0.2\ 0]^T$ ($t = 0$s) to $[0.2\ 0]^T$ ($t = 50$s). In particular, $x_d = [-0.2\cos(\frac{\pi}{50}t)\ 0.2\sin(\frac{\pi}{50}t)]^T$ is generated by Eq. (20) with

$$F(x_d) = \begin{bmatrix} 0 & \frac{\pi}{50} \\ -\frac{\pi}{50} & 0 \end{bmatrix} x_d \quad (73)$$

The equilibrium position of the human arm is $x_H = [0\ 0]^T$. The initial position of the robot arm in the joint space is $[-\frac{\pi}{2}\ \pi]^T$.

The NN parameters are: $l_c = l_a = l_{id} = 10$, $\mu_{c,ij} = \mu_{a,ij} = \mu_{id,ij} = -1 + 0.2i$, $\eta_{c,i} = 15$, $\eta_{a,i} = 10$, and $\eta_{id,i} = 100$ for $i = 1, 2, \ldots, l_c$ and $j = 1$, $j = 1, 2$, or $j = 1, 2, \ldots, 6$. The initial values are: $\hat{W}_c(0) = \mathbf{0}_{10}$, $\hat{W}_a(0) = \mathbf{0}_{10\times 2}$, and $\hat{W}_{id}(0) = \mathbf{0}_{10\times 6}$. The learning rates in the update laws (36), (38) and (54) are: $\sigma_c = \sigma_a = \sigma_{id} = 10$, and the control parameters $K_1 = I_2$, $K_2 = 600I_2$, $k_\Gamma = 3.5$, and $K_{id} = 10I_4$. A sweeping frequency signal $0.01\sum_{\varpi=1}^{10}\sin(\varpi t)$ is added into u to satisfy the condition of persistent excitation. This signal is small enough to not cause any disturbance to the human as it is smaller than the physiological tremor level.

A. Reinforcement Learning: Different Robots and Humans

To show the versatility of the proposed reinforcement learning control, we consider two different robot models and two human models, named Robot 1, Robot 2, Human 1

and Human 2. The parameters of Robot 1 are given as: $m_i = 10.00$kg, $l_i = 0.30$m, and $I_{zi} = 0.225$kgm^2, where m_i, l_i, and I_{zi}, $i = 1, 2$, represent the mass, the length, and the moment of inertia about the z-axis that comes out of the page passing through the center of mass, respectively. For Robot 2, $m_i = 20.00$kg and other parameters are the same to those of Robot 1. The impedance parameters C_H and K_H of Human 1 are determined from [42]: (21 ± 20)Ns/m and (201 ± 200)N/m, respectively. According to Assumption 1, we set $C_H = \text{diag}\{[21 - 20\cos(\dot{x}_1), 21 - 20\cos(\dot{x}_2)]\}$ and $K_H = \text{diag}\{[201 - 200\cos(\dot{x}_1), 201 - 200\cos(\dot{x}_2)]\}$, where $\dot{x}_1$ and $\dot{x}_2$ are two elements of the velocity $\dot{x}$. For Human 2, the impedance parameters are: $C_H = \mathbf{0}_2$ and $K_H = \text{diag}\{[100.5 - 100\cos(\dot{x}_1), 100.5 - 100\cos(\dot{x}_2)]\}$.

As discussed in Remark 1, different values of Q_1, Q_2, Q_3, and R can be chosen to penalize control objectives of human and robot. In the first case, we set $Q_1 = 100I_2$, $Q_2 = 0.01I_2$, $Q_3 = 0.001I_2$, and $R = 0.001I_2$ which indicate that robot's objective is more anticipated ($Q_1 = 100I_2$ and $Q_3 = 0.001I_2$).

The tracking performance for different combinations of robot and human models is illustrated in Figure 2. It is seen that although different models are adopted, the actual trajectory tracks the desired trajectory of the robot arm under the same control. This illustrates the learning capability of the proposed method. Norms of approximated critic and actor weights and approximated value functions in this case are shown in Fig. 3, which converge to different values for different combinations. Identification errors and control errors are shown to converge to small constants in Fig. 3.

In the second case, we set $Q_1 = I_2$, $Q_2 = 0.01I_2$, $Q_3 = I_2$, and $R = 0.05I_2$ which indicates that human's objective is more anticipated compared to the first case ($Q_1 = I_2$ and $Q_3 = I_2$). Tracking and approximation performances are illustrated in Figs. 4 and 5, respectively. From Fig. 4, we find that due to the effect of human control, the actual trajectory of robot arm drifts away from the desired one. For different combinations of robot and human models, effects of human control are different, but the convergence of the learning is always guaranteed, as shown in Fig. 5.

B. Comparison: Three Cases

To further show that control objectives of both human and robot can be reflected by choosing different values of Q_1, Q_2, Q_3, and R in Eq. (23), we consider the following criterion: if the human force f is larger than a threshold (0.1N in this simulation), it indicates that human wants to lead the task so $Q_1 = I_2$, $Q_2 = 0.01I_2$, $Q_3 = I_2$, and $R = 0.005I_2$; otherwise, $Q_1 = 100I_2$, $Q_2 = 0.01I_2$, $Q_3 = 0.001I_2$, and $R = 0.0005I_2$. Note that designing this criterion is task-dependent, while other options can be considered according to specific task objectives. We refer to the method based on the above criterion as "adaptive". For comparison, we consider another two cases with fixed weights:

- "human leading" with $Q_1 = I_2$, $Q_2 = 0.01I_2$, $Q_3 = I_2$, and $R = 0.005I_2$; and
- "robot leading" with $Q_1 = 100I_2$, $Q_2 = 0.01I_2$, $Q_3 = 0.001I_2$, and $R = 0.0005I_2$.

Different from the simulation in the previous subsection, the human intervention period is from $t_1 = \frac{50}{3}$s to $t_2 = \frac{100}{3}$s, i.e., $f = [0\ 0]^T$ for $t < t_1$ and $t > t_2$. Besides, models of Robot 1 and Human 1 are used in this simulation.

Actual trajectories of robot arm for three cases are shown in Fig. 6. Three phases are divided by $t = t_1$ and $t = t_2$. At the beginning of "adaptive" case, the actual trajectory tracks the desired trajectory of the robot arm after the learning period. When the human force is applied after $t = t_1$, the actual trajectory drifts away from the desired trajectory. When the human force disappears after $t = t_2$, the actual trajectory re-tracks the desired trajectory. These results are coherent with expectations. When human does not apply a force to the robot arm, Q_1 is given a relatively large value and Q_3 a small one, so robot's objective of trajectory tracking is more anticipated. When human applies a force larger than the prescribed threshold, he/she wants to lead the task, and Q_1 is given a relatively smaller value and Q_3 a larger one. As a result, the actual trajectory of the robot arm drifts away from the desired trajectory of the robot arm and to the equilibrium position of the human arm.

For "human leading" case, although the performance in the second phase is similar to that for "adaptive" case, the desired performance of trajectory tracking cannot be guaranteed in the other two phases.

Comparatively, for "robot leading" case, the desired performance of trajectory tracking is guaranteed in the first and third phases, but the robot arm cannot be moved to human's areas of interest. Results of tracking errors in the upper subfigure of Fig. 7 further confirm above discussions.

From the below subfigure of Fig. 7, we also find that a larger force will be resulted for "robot leading" case compared to the other two cases.

These results indicate that for either "human leading" or "robot leading" case, the following expected performance cannot be simultaneously achieved: a small tracking error when there is no human intervention and a small interaction force when there is human intervention. It can be only achieved for "adaptive" case. This "adaptive" case becomes feasible to realize with the proposed framework, where the weights in the novel cost functional can be modulated according to various situations and reinforcement learning guarantees the cost functional is minimized.

V. CONCLUSIONS

A framework of shared control via physical interaction has been designed in this work, with control objectives of both human and robot taken into consideration in a defined cost functional. Reinforcement learning has been employed to develop a control to minimize this cost functional in presence of unknown human and robot dynamics. Simulation results have been presented to show the learning capability of the proposed method and its feasibility in handling various situations.

One direction of our future works will be on specifying the defined cost functional according to the task requirements. Another direction will be on implementation of the proposed

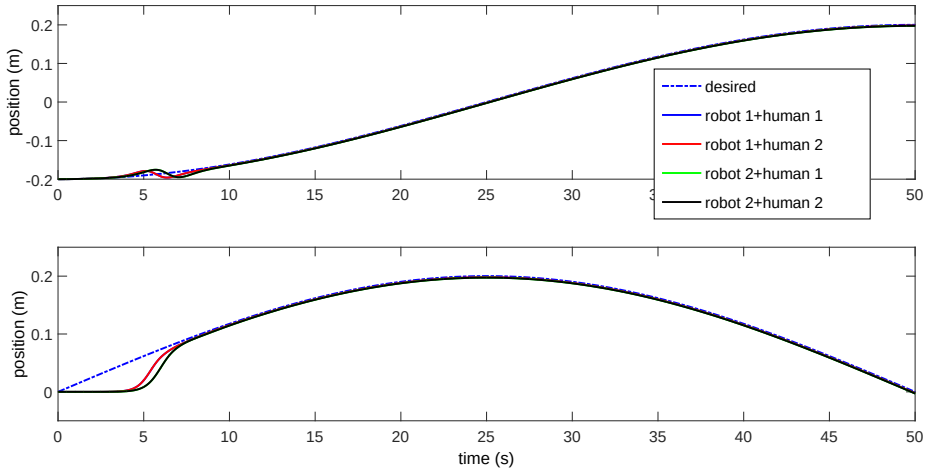

Fig. 2. The first case: actual trajectory and desired trajectory of robot arm for different combinations of robot and human models in X direction (upper) and Y direction (below)

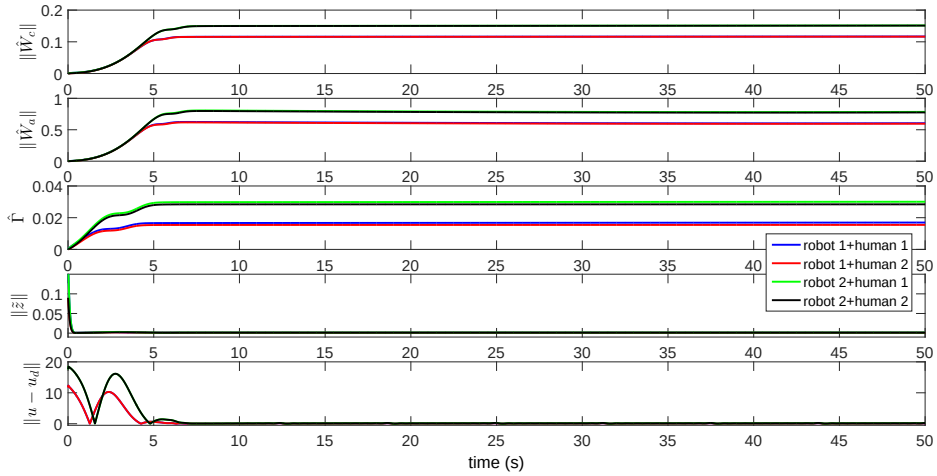

Fig. 3. The first case: norms of approximated critic weights, norms of approximated actor weights, approximated value functions, identification errors, and control errors for different combinations of robot and human models

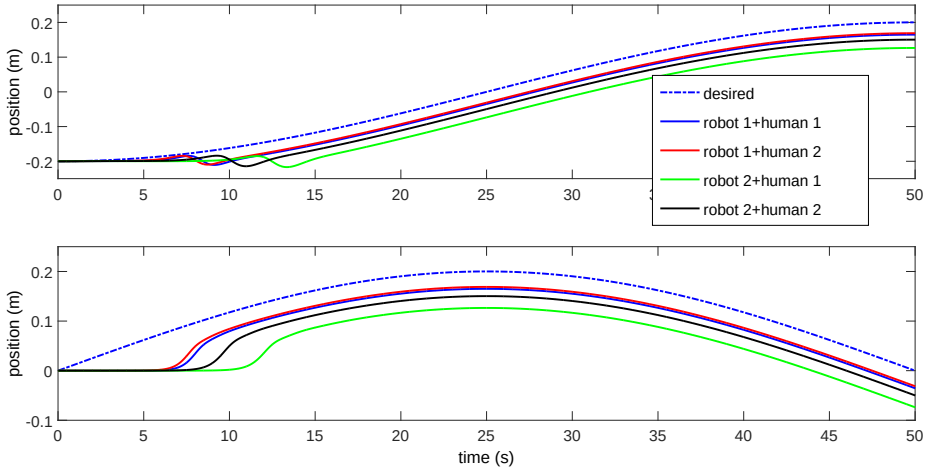

Fig. 4. The second case: actual trajectory and desired trajectory of robot arm for different combinations of robot and human models in X direction (upper) and Y direction (below)

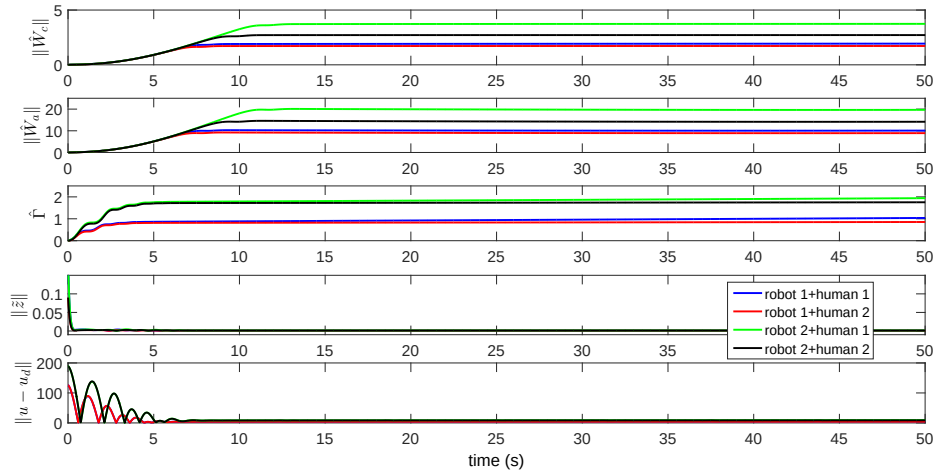

Fig. 5. The second case: norms of approximated critic weights, norms of approximated actor weights, approximated value functions, identification errors, and control errors for different combinations of robot and human models

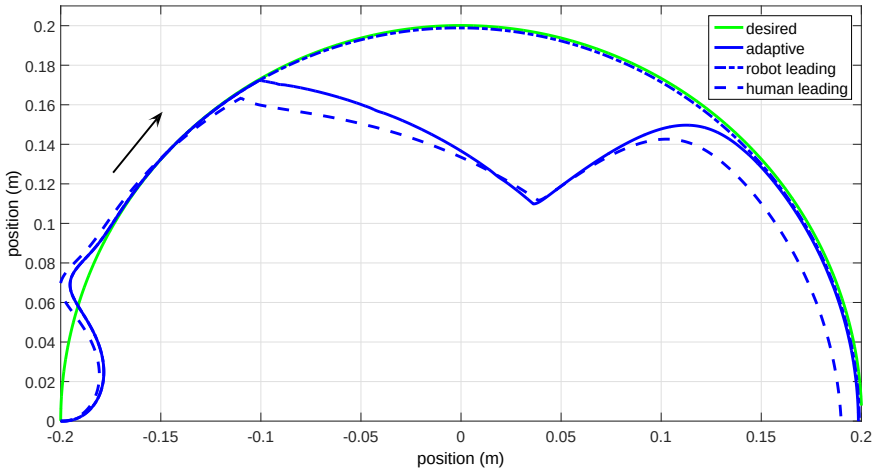

Fig. 6. Actual trajectories for 1) "adaptive" case, 2) "human leading" case, and 3) "robot leading" case, and desired trajectory of robot arm. The arrow indicates the motion direction of the end-effector of the robot arm.

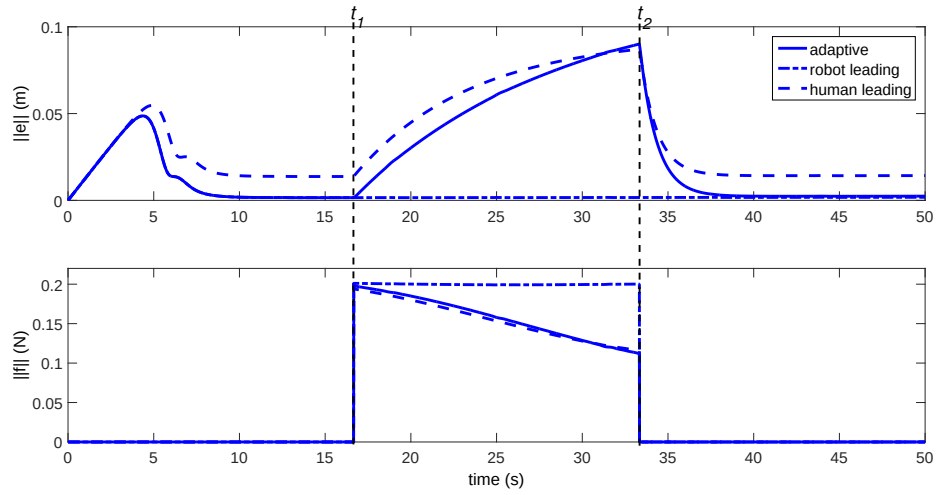

Fig. 7. Norms of tracking errors (upper) and interaction forces (below) for 1) "adaptive" case, 2) "human leading" case, and 3) "robot leading" case. t_1 and t_2 are the start time and end time of human intervention, respectively.

framework on physical robots. Finally, human behaviours may not be completely described by the model (6), so how they will affect the performance of the proposed method will be studied in a real-world application.

REFERENCES

[1] X. Wang, C. Yang, H. Ma, and L. Cheng, "Shared control for teleoperation enhanced by autonomous obstacle avoidance of robot manipulator," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4575–4580, Sept 2015.

[2] W. He, Z. Li, and C. L. P. Chen, "A survey of human-centered intelligent robots: issues and challenges," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 602–609, 2017.

[3] C. Zeng, C. Yang, Z. Chen, and S.-L. Dai, "Robot learning human stiffness regulation for hybrid manufacture," *Assembly Automation*, 2018.

[4] C.-H. Ko, Y.-C. H. K.-Y. Young, and S. K. Agrawal, "Walk-assist robot: A novel approach to gain selection of a braking controller using differential flatness," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 6, pp. 2299–2305, 2013.

[5] M. S. Erden and B. Maric, "Assisting manual welding with robot," *Robotics and Computer-Integrated Manufacturing*, vol. 27, no. 4, pp. 818–828, 2011.

[6] J. A. Shah, J. H. Saleh, and J. A. Hoffman, "Review and synthesis of considerations in architecting heterogeneous teams of humans and robots for optimal space exploration," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 5, pp. 779–793, 2007.

[7] N. Hogan, "Impedance control: an approach to manipulation-Part I: Theory; Part II: Implementation; Part III: Applications," *Transaction ASME J. Dynamic Systems, Measurement and Control*, vol. 107, no. 1, pp. 1–24, 1985.

[8] Q. Liu, C. Shi, B. Zhang, C. Wang, L. Duan, T. Sun, X. Zhang, W. Li, Z. Wu, and M. G. Fujie, "Development of a novel paediatric surgical assist robot for tissue manipulation in a narrow workspace," *Assembly Automation*, vol. 37, no. 3, pp. 335–348, 2017.

[9] N. Jarrasse, V. Sanguineti, and E. Burdet, "Slaves no longer: review on role assignment for human-robot joint motor action," *Adaptive Behavior*, vol. 22, no. 1, pp. 70–82, 2014.

[10] J. R. Medina, D. Lee, and S. Hirche, "Risk-sensitive optimal feedback control for haptic assistance," in *IEEE International Conference on Robotics and Automation*, pp. 1025–1031, 2012.

[11] J. R. Medina, T. Lorenz, D. Lee, and S. Hirche, "Disagreement-aware physical assistance through risk-sensitive optimal feedback control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3639–3645, 2012.

[12] J. R. Medina, M. Lawitzky, A. Molin, and S. Hirche, "Dynamic strategy selection for physical robotic assistance in partially known tasks," in *IEEE International Conference on Robotics and Automation*, pp. 1180–1186, 2013.

[13] M. S. Erden and T. Tomiyama, "Human-intent detection and physically interactive control of a robot without force sensors," *IEEE Transactions on Robotics*, vol. 26, no. 2, pp. 370–382, 2010.

[14] K. Wakita, J. Huang, P. Di, K. Sekiyama, and T. Fukuda, "Human-walking-intention-based motion control of an omnidirectional-type cane robot," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 1, pp. 285–296, 2013.

[15] Y. Li and S. S. Ge, "Human-robot collaboration based on motion intention estimation," *IEEE/ASME Transactions on Mechatronics*, vol. 19, no. 3, pp. 1007–1014, 2014.

[16] M. S. Erden and A. Billard, "End-point impedance measurements at human hand during interactive manual welding with robot," *Proceedings of IEEE International Conference on Robotics & Automation*, pp. 126–133, 2014.

[17] R. Chipalkatty, G. Droge, and M. B. Egerstedt, "Less is more: Mixed-initiative model-predictive control with human inputs," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 695–703, 2013.

[18] P. J. Werbos, "A menu of designs for reinforcement learning over time," *Neural Networks for Control*, pp. 67–95, 1990.

[19] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[20] D. Huang, J.-X. Xu, V. Venkataramanan, and T. Huynh, "High-performance tracking of piezoelectric positioning stage using current-cycle iterative learning control with gain scheduling," *IEEE Transactions on Industrial Electronics*, vol. 61, pp. 1085–1098, Feb 2014.

[21] D. Huang, J.-X. Xu, S. Yang, and X. Jin, "Observer based repetitive learning control for a class of nonlinear systems with non-parametric uncertainties," *International Journal of Robust and Nonlinear Control*, vol. 25, no. 8, pp. 1214–1229, 2015.

[22] Z. Chen, C. Yang, X. Liu, and M. Wang, "Learning control of flexible manipulator with unknown dynamics," *Assembly Automation*, vol. 37, no. 3, pp. 304–313, 2017.

[23] W. He, T. Meng, X. He, and S. S. Ge, "Unified iterative learning control for flexible structures with input constraints," *Automatica*, vol. 96, pp. 326 – 336, 2018.

[24] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers," *IEEE Circuits and Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[25] D. P. Bertsekas, M. L. Homer, D. A. Logan, S. D. Patek, and N. R. Sandell, "Missile defense and interceptor allocation by neuro-dynamic programming," *IEEE Transactions on System, Man, and Cybernetics, Part A*, vol. 30, no. 1, pp. 42–51, 2000.

[26] Y. Jiang and Z. P. Jiang, "Robust adaptive dynamic programming for large-scale systems with an application to multimachine power systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 10, pp. 693–697, 2012.

[27] Y. Hirata, Z. Wang, K. Fukaya, and K. Kosuge, "Transporting an object by a passive mobile robot with servo brakes in cooperation with a human," *Advanced Robotics*, vol. 23, pp. 387–404, 2009.

[28] M. W. Spong and M. Vidyasagar, *Robot Dynamics and Control*. New York, USA: Wiley, 1989.

[29] T. Tsuji, P. G. Morasso, K. Goto, and K. Ito, "Human hand impedance characteristics during maintained posture," *Biological Cybernetics*, vol. 72, no. 6, pp. 475–485, 1995.

[30] V. Duchaine and C. Gosselin, "Safe, stable and intuitive control for physical human-robot interaction," in *IEEE International Conference on Robotics and Automation*, pp. 3676–3681, 2009.

[31] K. Doya, "Reinforcement learning in continuous time and space," *Neural Computation*, vol. 12, no. 1, pp. 219–245, 2000.

[32] W. He and Y. Dong, "Adaptive fuzzy neural network control for a constrained robot using impedance learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 1174–1186, April 2018.

[33] R. M. Sanner and J. E. Slotine, "Gaussian networks for direct adaptive control," *IEEE Transactions on Neural Networks*, vol. 3, no. 6, pp. 837–863, 1992.

[34] D. Vrabie and F. L. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, 2009.

[35] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, 2014.

[36] K. S. Narendra and A. M. Annaswamy, "Persistent excitation in adaptive systems," *International Journal of Control*, vol. 45, no. 1, pp. 127–160, 1987.

[37] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[38] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[39] P. He and S. Jagannathan, "Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints," *IEEE Transactions on System, Man and Cybernetics, Part B*, vol. 37, no. 2, pp. 425–436, 2007.

[40] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning with explorations for continuous-time nonlinear systems," in *International Joint Conference on Neural Networks*, pp. 1–6, June 2012.

[41] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–1, 2014.

[42] S. P. Buerger and N. Hogan, "Complementary stability and loop shaping for improved human-robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 2, pp. 232–244, 2007.